Parameter selection for segregating speech from background noise

Honors Undergraduate Research Thesis

Presented in partial fulfillment of the requirements for graduation with *honors research distinction* in Speech and Hearing Science in the undergraduate colleges of The Ohio State University

by
Jordan Vasko

The Ohio State University
May 2015

Project Advisor: Dr. Eric W. Healy, Department of Speech and Hearing Science

Table of Contents

Abstract

Understanding speech in background noise remains a primary challenge faced by hearing-impaired listeners. Ideal binary masking (IBM) is an effective technique to facilitate understanding of a target signal in noisy backgrounds, and IBM estimation is the goal of an effective speech-from-noise separation algorithm that holds promise for alleviating limitations of hearing impairment. In IBM processing, a speech-and-noise mixture is divided into a grid of time-frequency (T-F) units, which are discarded if their degree of noise corruption (reflected as a signal-to-noise ratio, or SNR) exceeds a certain local criterion (LC). Prior work determined that the relationship between the overall SNR of the original speech-noise mixture and LC (the relative criterion or RC) was important for determining intelligibility. This prior work also suggests that there is a wide range of RC values over which performance scores reach maximum. The current study investigates whether these scores reflect a performance ceiling rather than a true maximum. Consonant recognition was tested in normal-hearing listeners using seven different RC values. The background was speech-shaped noise. An RC performance function was obtained that did not display the ceiling effect limitations of previous work. This function suggests that the optimal RC value may be different from previous estimates. These findings have implications for selections of LC during IBM estimations. They also suggest appropriate parameters for testing the effect of varying LC within a single mask according to specific frequency contributions to overall speech intelligibility. Such developments may contribute to reducing the struggles that hearing-impaired listeners face in noise.

I. Introduction

The single largest deficit that accompanies hearing loss is the inability to listen to a target signal (e.g., speech) when it is embedded in background noise (Dillon, 2012). Current hearing aids primarily amplify incoming sounds. Although most have some noise reduction capabilities, they still have limited ability to segregate a target signal from background noise. Therefore, many of the 360 million hearing-impaired individuals worldwide (World Health Organization, 2014) suffer from difficulties in speech understanding in noisy environments, even while using hearing aids. This problem is significant enough that its solution has been called the “holy grail” of the field.

One particularly effective, though not directly implementable, approach to helping hearing-impaired listeners segregate speech from noise is referred to as the ideal binary mask (IBM). In IBM processing, a sound signal is parceled into a grid of time-frequency (T-F) units reflecting the specific time window and frequency band that describe the unit. Prior knowledge of the acoustic characteristics of both the target speech signal and the background noise (that is, both signals are labeled and fed into the processing script separately) is used to identify speech-dominated versus noise-dominated units. Speech-dominated units are retained and noise-dominated units are discarded before the ideal-binary masked signal is transmitted to a listener.

The decision to retain or discard a T-F unit is dependent on whether that unit exceeds a predetermined signal-to-noise ratio (SNR) of the intensity of the target signal compared to that of the background signal (Hu & Wang, 2001; Wang, 2005). The SNR value on which this decision is made is referred to as the local criterion (LC). Mathematically, this processing can be described by the simple step function:

$$IBM(t,f) = \begin{cases} 1, \text{if } SNR\ (t,f) > LC \\ 0, \text{otherwise,} \end{cases} \quad (1)$$

Where mask values of 1 indicate retaining of the unit, and mask values of 0 indicate discarding of the unit. Note that the indexation *(t,f)* references either the mask value (*IBM*) or local SNR (*SNR*) of a specific time-frequency unit.

Figure 1, reprinted from Wang et al. (2008), is a visualization of IBM processing. The top panels of the figure represent 32-channel cochleagrams of a sentence spoken in quiet and of an SSN masker. The bottom left panel is the visualization of the ideal binary mask, where the black space represents the T-F units that were discarded because the SNR was less than the local criterion, and the white space represents units that were classified as speech-dominant and retained. The bottom right panel represents the speech-and-noise mixture that has been processed through the IBM. Much of the signal has been discarded, but the vowel formants and consonant onsets are still noticeable.

The use of a *local* criterion is beneficial because the level of the speech fluctuates across the frequency spectrum throughout the duration of the masking signal. Therefore, whether the background noise is steady-state (e.g. speech-shaped noise) or is also modulated, there will be points during the duration of the mixed signal for which, in at least some frequency bands, the local SNR is greater than the overall SNR. Even for relatively low overall SNRs, there may be some time-frequency units that are relatively "clean" and free of noise. Selecting a high LC value will yield a comparatively selective mask for which only very few, very clean T-F units remain, while selecting a low LC value may preserve too many noisy T-F units to improve intelligibility of the processed signal.

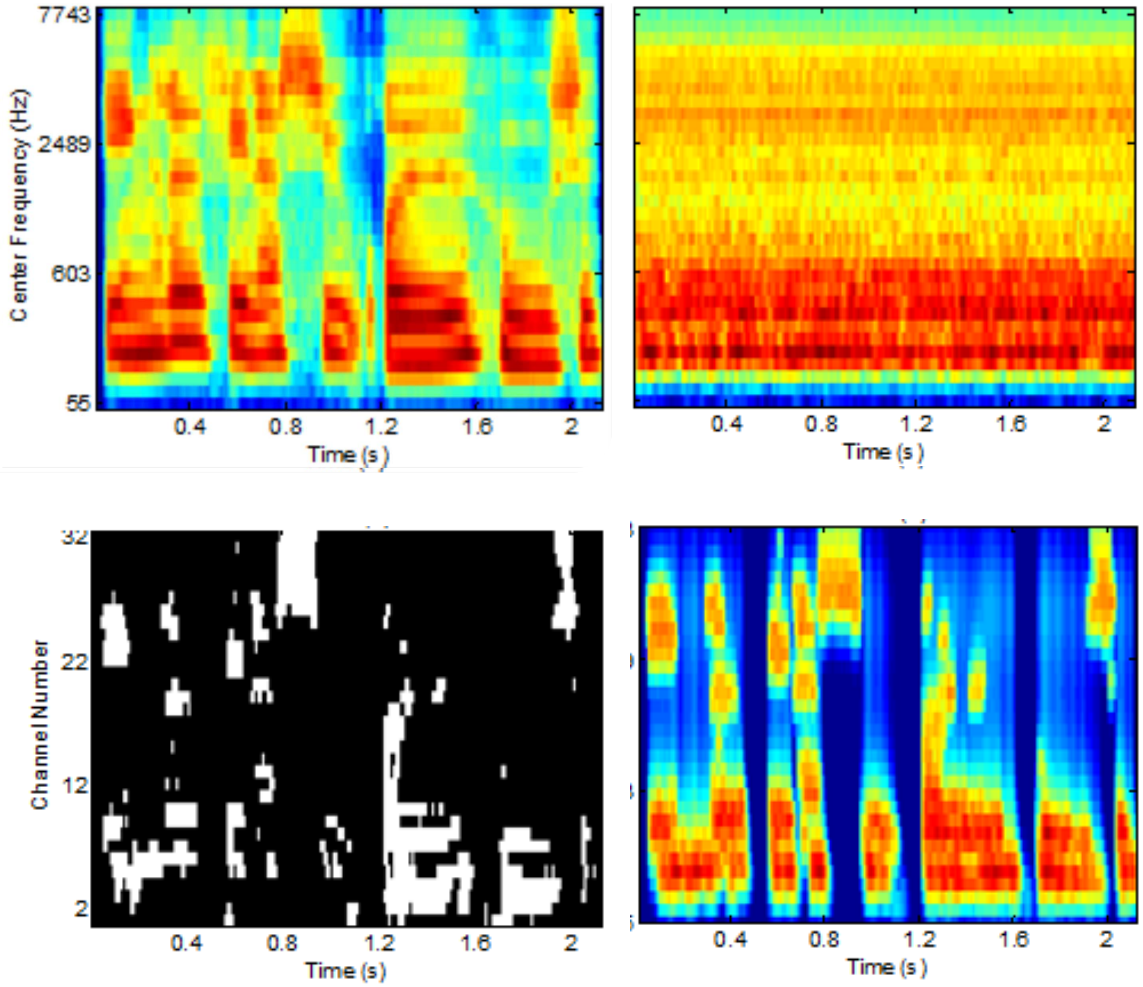

FIGURE 1. 32-channel cochleagrams of normal speech, speech-shaped noise, ideal binary mask (IBM), and speech-noise mixture after IBM is applied. From Wang et al. (2008).

The IBM has been shown to be extremely successful, as hearing-impaired listeners tested in different noise types can obtain speech understanding scores nearly equivalent to those of normal-hearing listeners (Wang et al., 2009). Explanations for IBM effectiveness suggest that the mask facilitates the glimpsing of speech in noise, whereby "glimpses" of spectrotemporal portions of a speech signal through a noise background facilitate segregation of the two signals (Cooke, 2006). The pattern of the IBM, therefore, is responsible for intelligibility improvements because it cues listeners as to which units are the clean glimpses.

The tradeoff between changing the LC value of the IBM and changing the overall SNR of the original, unprocessed mixture was first discussed by Brungart et al. (2006), in which it was emphasized that it is not the LC itself, but rather the difference between the LC and the input SNR that is important in determining the effectiveness of the IBM. Kjems et al. (2009) defined this difference as a relative criterion (RC). The RC of a mask can be described as follows:

$$RC = LC - \text{Overall (Input) SNR} \qquad (2)$$

It was found that the same mask will be produced (with respect to the retention and discarding of the same T-F units) if a stimulus is processed with two different initial overall SNRs but the same RC value (Brungart et al., 2006). Therefore similar RC values across speech-noise mixtures of varying SNRs and speech material also represent similar proportions of a mask being retained and discarded.

Brungart et al. (2006) processed two sets of stimuli through the IBM such that one set had a constant original mixture SNR but varied in terms of LC value, and the other set used a constant LC value but varied in terms of original mixture SNR. Thus the two sets had the same

ranges of RC values, but from different sources. It was shown that performance was nearly identical when compared across these two sets of stimuli.

Kjems et al. (2009) found a performance function for normal-hearing listeners across different RC values. This study provided further evidence that speech recognition performance can be very similar for stimuli with very different input SNRs if the same RC is used in processing. Indeed, when normal-hearing listener performance was plotted against RC, the function was very similar for a wide range of overall SNRs, and this result held across a variety of noise types.

Important for the current work is the fact that Brungart et al. (2006) and Kjems et al. (2009) revealed ceiling effects in which performance neared 100% correct for a wide range of RC values. These studies employed low-predictability sentences (the CRM and Dantale II corpora, respectively) in multitalker babble and speech-shaped noise (SSN), respectively. Brungart et al. (2006) was not able to alleviate the ceiling effect in any noise condition. However, they still propose a performance plateau from -12 to 0 dB SNR (from an overall SNR of 0 dB SNR and for between two and four competing talkers). Kjems et al. (2009) employed an extremely negative overall SNR (-60 dB) to yield scores far enough below the performance ceiling to estimate optimal RC values. At such a low SNR, virtually no speech information was available in the unprocessed mixture. The optimal RC values reported there ranged from -8.8 to -1.6 dB SNR. This range corresponds to retention of approximately 20% to 40% of the original mixture.

The center of this performance plateau is estimated to be -5 dB SNR in an SSN masker. Whether there are statistically significant differences between any of the conditions representing

adjacent RC values remains unclear, as no post-hoc analysis of these differences was mentioned in the earlier publications. The current experiment attempts to replicate and extend those experiments. A performance ceiling is avoided by having listeners complete a more difficult experimental task at a more ecologically valid SNR, thus better approximating a true optimum RC to perhaps use in future algorithm processing.

II. Method

The aim of the current experiment was to determine the RC performance function (the performance based on a number of different RCs) for the IBM when performance scores were lowered below the 100% ceiling found in other publications. Normal-hearing listeners were recruited to follow the convention employed in prior literature and to establish a benchmark for the function in the normal auditory system. Furthermore, the variability and scarcity of hearing-impaired listeners implies that a more reliable function would be calculated from normal-hearing listeners.

Subjects

Ten participants were recruited from undergraduate Speech and Hearing Science courses, and were compensated with extra course credit. All were native English speakers with normal hearing, defined by hearing thresholds less than or equal to 20dB HL in the frequency range 250 Hz through 8000 Hz. A hearing screening was performed by the experimenter before each subject was run to confirm normal hearing. All participants were female. The average age of participants was 20.3 yrs., and ranged from 20 to 22 yrs.

Stimuli

The stimuli were created using recordings from four male talkers uttering consonant phonemes in an /aCa/ environment. Though members of a closed set, the /aCa/ phonemes provide almost no semantic cues, and therefore were hypothesized to lead to lower performance scores (and thus to avoid a ceiling effect). Scores reported in Simpson and Cooke (2005) and Healy et al. (2014), also suggest that the use of /aCa/ phoneme stimuli at an SNR comparable to that used currently should decrease performance scores to as low as 50% correct. The following

16 consonants were used (as in Healy et al., 2014): /p, t, k, b, d, g, f, v, s, z, ʃ, ʒ, θ, ð, m, n/. Recordings were made using four different male talkers.

The individual /aCa/ consonant recordings were normalized in terms of total RMS power before any processing occurred. However, because more T-F units are discarded in conditions with higher RC values, the overall loudness and total RMS power of the all of the processed stimuli presented in the experiment were not equivalent.

The consonant recordings were mixed with steady-state speech-shaped noise (SSN). The SSN was generated in MATLAB from white noise (which was generated in Cool Edit to match the duration of the concatenated set of 64 /aCa/ phonemes) using a 1,000-order arbitrary-response finite impulse-response filter (fir2) with a Hamming-windowed fast-Fourier transform. The frequency spectrum and amplitude of the SSN was generated to match the long-term average spectrum of the 64 recorded consonants played in quiet. The frequency and amplitude characteristics were calculated using SpectraPlus software.

IBM processing in this series of experiments was nearly identical to that used in Healy et al., 2014, with the exception of specific LC and input SNR values used. Sampling occurred at a frequency of 44.1 kHz with a 16-bit depth. Synthesis of all speech, noise, and mixed signals occurred in MATLAB, using the IBM script produced by D. L. Wang, Z. Z. Jin, Y. P. Li, and J. F. Woodruff (found on D. L. Wang's website at http://web.cse.ohio-state.edu/pnl/shareware/cochleagram/), and modified by F. Apoux and the author of this document. Time windows were 20 ms in duration with a 10 ms overlap between adjacent windows.

Speech and noise were filtered separately by 64 gammatone filters (each one equivalent rectangular bandwidth [ERB] in width) before the computation of the IBM. Filtering was applied to the same frequency range as in Kjems et al. (2009) (center frequencies of each ERB were between 50 and 8000Hz) to facilitate comparisons.

Seven conditions were employed that differed only in terms of RC value. The overall SNR in each condition was -8 dB. The LC was varied from -28 dB SNR to +2 dB SNR in increments of 5 dB, to thus test an RC range of -20 dB SNR to +10 dB SNR and 7 different RC values total. Lower RC values reflect IBMs in which a greater proportion of the original mixture is retained, and higher RC values reflect IBMs in which a greater proportion of the original speech and noise are discarded. The percentages of the original mixture retained range from approximately 80% at RC = -20 dB SNR to approximately 2% at RC = 10 dB SNR. See Table 1 for a summary of the conditions employed.

Table 1. Experimental Conditions

LC (dB SNR)	Input SNR (dB SNR)	RC (dB SNR) [=LC - Input SNR]	Percent Ones*
-28	-8	**-20**	79.83%
-23	-8	**-15**	69.10%
-18	-8	**-10**	55.65%
-13	-8	**-5**	37.07%
-8	-8	**0**	17.00%
-3	-8	**5**	7.24%
2	-8	**10**	2.56%

*in the IBM; averaged across the Talker 1 masks for all 16 consonants

The aforementioned range of RC values was selected because it encompassed the RC values for which performance scores were at the 100% correct ceiling for input SNRs similar to those used currently as well as values approximately 10 dB SNR above and below the proposed performance plateau in Kjems et al. (2009). Results from Healy et al. (2014) suggest that consonant stimuli in SSN at the SNR and RC values used in the current experiment should lead to performance scores by normal-hearing listeners near a maximum of 71%.

Procedure

Presentation of stimuli occurred within a double-walled sound-treated booth via supra-aural Sennheiser HD 200 Pro headphones. Responses were forced-choice with 16 alternatives. After each trial, participants used the mouse to select the phoneme heard from a grid of letters on a computer monitor. It was made clear before the experiment which letter represented which phoneme, and participants were given a reference sheet for use during the experiment describing the sound of each phoneme. The reference sheet is included as Figure 2. The reference sheet closely resembled the display on the monitor, but also included example words containing the phoneme. Responses were scored in terms of percent of consonants identified correctly.

Participants were required to complete a practice session before the experimental task. During the practice session, participants identified all of the /aCa/ phonemes (in random order) in quiet and without IBM processing, using the same interface that they would during the experiment. Feedback was provided after each response by visual identification of the phoneme actually presented during the trial.

The practice was intended to familiarize participants with the voices of each of the four talkers, and as a confirmation that each participant could perform the task. Participants were

required to attain a score of at least 90% before continuing with the experiment. Participants who did not earn this score after their first practice session were trained on the differences between the specific consonants that caused their poor performance and then presented with the practice task again. No subject completed the practice more than two times. One subject earned only 89% in her second practice session, but was given further training on the two consonants on which she made consistent errors before she began the experimental task.

P pine	T time	K kite	B bite
D dime	G guide	F fight	S sew
Sh shoe	V vice	Z zebra	Dz measure garage
M mom	N nice	Th thin bath	Thz these bathe

FIGURE 2. The reference sheet given to participants during the experimental task to help illustrate the specific phonemes represented by each letter (or group of letters). The display on the computer monitor closely resembled this reference sheet, but did not include the example words under each letter.

Furthermore, in preparation for the experiment, subjects were introduced to the sound quality of IBM-processed speech. The consonants as produced by Talker 1 were selected from the RC = -5 dB SNR condition and played to participants, who were made aware of which consonant was contained within each presentation. It was assumed that because so few of the stimuli from this condition were presented, and presented so briefly, any effect that this familiarization stage would have on final performance scores would be negligible.

The level of presentation was calibrated such that the average RMS power of the unprocessed, quiet signal was 65 dBA as measured on a flat-plate coupler. Participants heard each of the 64 phonemes once per condition. Given the short duration (approximately 40 minutes) of each experiment, each subject completed all seven of the conditions in one sitting. Conditions were presented in blocks, with the order of blocks randomized for each participant. Order of phonemes was also randomized within each condition.

IV. Results

Table 2 shows the mean percent-correct scores and standard deviations obtained in each condition. Figure 3 shows the RC performance function for each individual listener in each condition, with RC value plotted against percent of consonants identified correctly. The bold line represents the mean score across listeners in each condition. The figure shows that no participant's score in any condition was greater than 90%, indicating that the influence of a ceiling effect had indeed been avoided.

Table 2. Means and standard deviation of percent-correct performance scores by RC value

RC value	Mean	Std Dev
-20	71.88	5.21
-15	80.47	6.00
-10	81.41	3.79
-5	78.13	4.54
0	68.28	4.66
5	62.81	3.59
10	51.25	8.71

A one-way repeated measures ANOVA was conducted on the current data, and revealed that a statistically significant difference existed amongst the seven conditions with $F(6,54) = 43.9$, $p < 0.001$ (significant at familywise $\alpha = .05$). Arcsine transform was unnecessary because of the range of scores obtained.

Figure 4 shows group mean consonant recognition scores plotted with standard error bars for each condition. This graph illustrates the existence of a plateau in optimal performance across

a modest range of RC values (from -15 to -5 dB SNR). No subject displayed best performance (nor one of her three best scores) in the RC = 0 dB SNR condition, which had been identified in Brungart et al. (2006) as part of the "region" of optimal RC values. Nine of the ten subjects achieved their highest three scores in the three conditions identified in the current performance plateau.

The results of a post hoc analysis are summarized in Table 3. Pairwise Bonferroni comparisons revealed that scores resulting from RC values within the plateau region were significantly different from those in adjacent conditions, with $p < .001$ for all statistically significant pairwise comparisons. Of note is that the RC = -5 dB SNR condition was not statistically different from the RC = 20 dB SNR condition ($t = 2.7$, $p = 0.010$ $\mu_{-5} - \mu_{-20} = 6.3$; $s_{-5} = 4.5$; $s_{-20} = 5.2$), even though the other RC values comprising the plateau (-15 dB and -10 dB) were both significantly different from RC = -20 dB SNR ($t = 3.7$ for -15 dB SNR, $t = 4.10$ for -10 dB SNR).

The power for the 21 comparisons conducted between scores for all pairs of RC values (pairwise $\alpha = 0.0024$) was expected to be approximately 0.85, which exceeds the conventionally desirable 80% power. Therefore it is reasonable to expect that a true difference that existed between means in adjacent conditions would have been revealed by this analysis, and that the proposed performance plateau is a real effect.

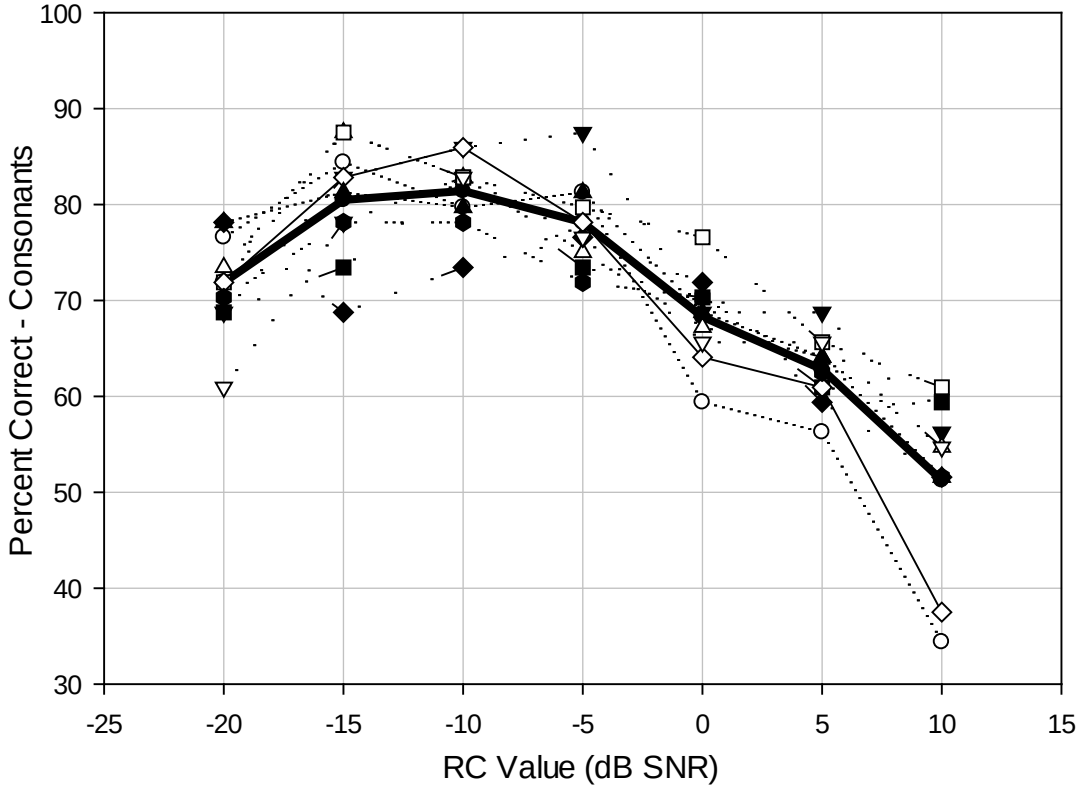

FIGURE 3. /aCa/ performance for each of the 10 listeners across conditions. Each black line represents a separate listener, and the bold line with filled circles represents the mean across subjects.

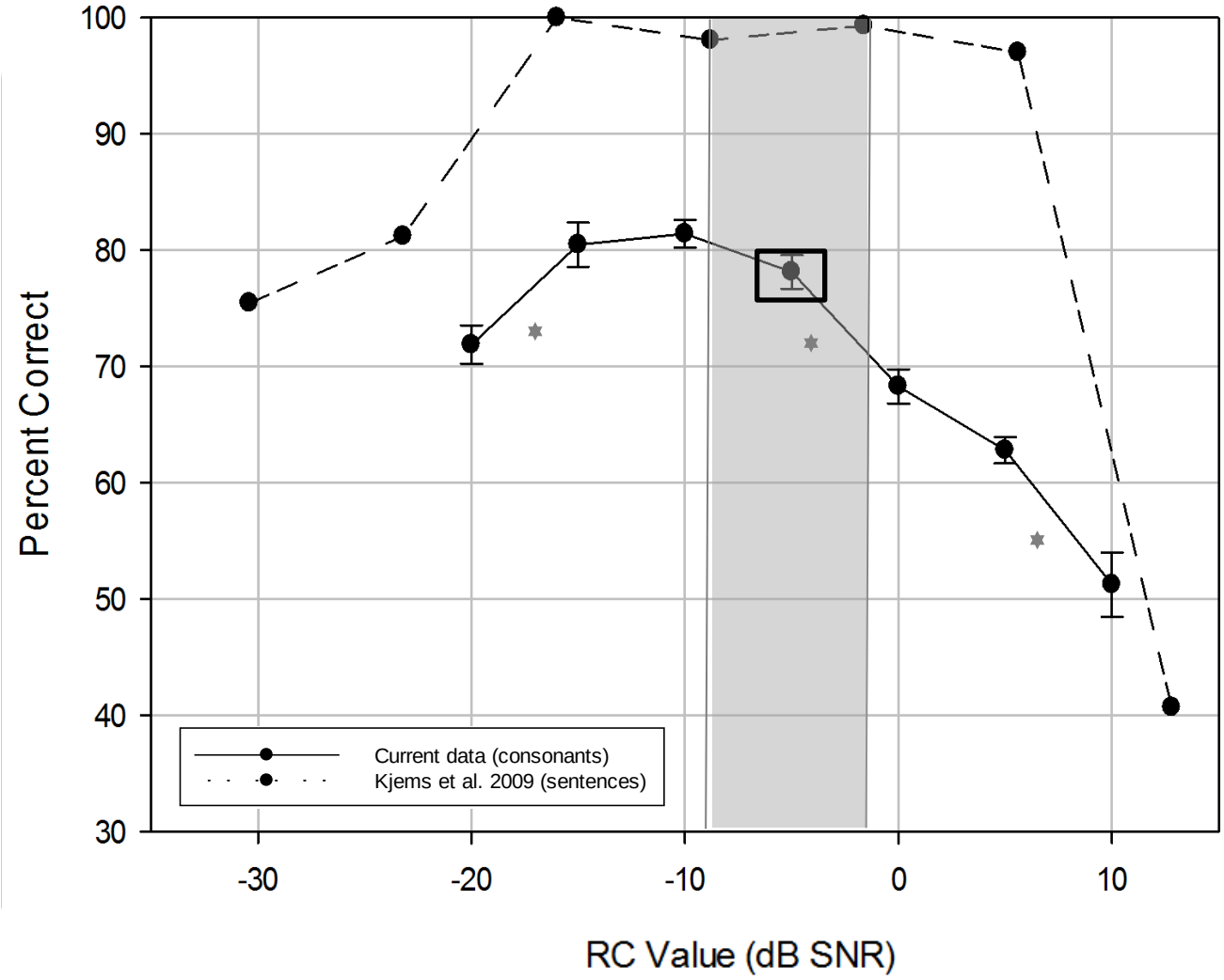

FIGURE 4. Mean RC performance function (with SEs) for the current study and for Kjems et al. (2009). The gray region is the plateau proposed by Kjems et al. For the current data, asterisks (*) indicate statistically significant differences between adjacent conditions. The box indicates the RC values used in the IBM estimation algorithms (e.g. Healy et al., 2013, 2014).

Table 3. Bonferroni post-hoc results

Condition	Significantly different from*	Not significantly different from*
-20	**-15,** -10, 5, 10	0, -5
-15	**-20**, 0, 5, 10	**-10, -5**
-10	-20, 0, 5, 10	**-15, -5**
-5	**0**, 5, 10	-20, **-15, -10**
0	-15, -10, **-5**, 10	-20, **5**
5	-20, -15, -10, -5, **10**	0
10	-20, -15, -10, -5, 0, **5**	n/a

*Pairwise comparisons are significantly different at $p < 0.0024$ for familywise $\alpha = .050$ with 21 comparisons. One-way RM ANOVA was significant with $F(6,54) = 43.9$, $p < 0.001$. Bolded values represent comparisons between adjacent conditions.

V. Discussion

An important difference from previous data was discovered in the current performance scores. Though the width of the performance plateau is approximately the same as had been estimated from previous experiments, the RC values responsible for this plateau have shifted. That is, current data show that maximum performance scores result from RC values between -15 to -5 dB SNR and appear to be centered around -10 dB SNR. Estimates show that between approximately 70% and 37% of the T-F units are retained for the RC values in the performance plateau. This reflects a benefit from the retention of more of the original speech-noise mixture compared to the results of Kjems et al. (2009), in which performance was optimized when approximately 20% to 40% of the original mixture was retained. More noise than previously estimated was tolerated by NH listeners in exchange for the provision of more speech information.

It is interesting that, even in the absence of a ceiling effect, there are several RC values that yield similar performance. This implies that human listeners perform equivalently across a range of relative criteria values and therefore across a wide variety of proportions of the original speech-noise mixture that are retained. Specifically, performance appears to be maximized and is not significantly different among conditions as long as between approximately one-third and two-thirds of the mixture is retained. This finding suggests that tradeoff between the benefits of discarding more speech and retaining more noise is equivalent over a wide range. Further investigation into this finding may yield information regarding noise tolerance in normal-hearing listeners.

There are several possible interpretations for the finding that the optimal RC values as identified in the current study are slightly more negative than previously indicated. The favoring of more signal retention may be interpreted as an effect of using a more ecologically valid signal-to-noise ratio (SNR) in the current research. Therefore the values included in the optimal range of RCs as suggested by the current experiment may be considered more reliable than previous estimates. It is possible that when the SNR is large enough that some speech remains in the mixture (and thus when the SNR is more ecologically valid), it is more beneficial to retain a greater proportion of the original mixture than when the SNR is so low as to reflect an unprocessed signal that is essentially noise. But it is also important to note that isolated consonants were employed currently, rather than simple sentences as employed previously.

Kjems et al. (2009) did not report post-hoc analyses regarding differences in participant performance according to RC value. However, visual inspection of the -60 dB SNR conditions in Figures 3, 4, and 7 of that publication suggests that for certain noise types, there may not be a performance plateau, but rather one RC value that leads to maximal performance. A visual examination of Figures 3 and 4 from Kjems et al. suggests that extremely unfavorable SNRs lead to optimal RC value estimates that are more positive than conditions with higher SNRs. This may be because the left portion of the performance function steepens as listeners are tested in increasingly negative SNRs, but the right portion of the function is relatively stable across various SNRs. Thus, it is possible that the RC performance function may be artificially narrow and shifted at extremely unfavorable (and unrealistic) SNRs. Such a finding may limit the assertion of previous research that listener performance is similar across IBM-processed stimuli with different original mixture SNRs but equivalent RC values.

However, potential effects of speech material must also be considered. Stimuli for previous experiments related to the effect of LC/RC value had been sentences low predictability but for which the words to be identified were of a closed-set and could be repeated across sentences (CRM corpus; Brungart, et al., 2006) or not (Dantale II sentences; Kjems et al., 2009). In both cases more linguistic and acoustic information was available (from knowledge of the language and coarticulation) than was available from the consonant stimuli used in the current experiment. The additional speech information available in the sentence stimuli may allow for intelligibility to be preserved when a greater proportion of the speech signal is discarded entirely. Therefore, more of both the signal and noise may need to be retained as cues are scarcer in the /aCa/ consonants.

VI. Conclusion and Future Directions

1. Variable-LC Mask

IBM effectiveness under different processing conditions (which include effects of LC, frequency channel resolution, and under different noise types, among others) is explained and supported by a large body of literature (Anzalone et al., 2006; Brungart et al., 2006; Li and Loizou, 2008; Wang et al., 2008, 2009; Cao et al., 2011; Sinex, 2013). However, there is no record of experiments that test the application of the different and inequivalent contributions of speech information of different frequency bands. The optimal range of RC values as identified in this paper may be used to specify parameters for a "variable-LC" mask in which the LC is varied within the same mask to account for the different importance of each band to speech understanding.

The concept that some frequency bands are more important for speech understanding than others has been long established in the field of hearing science (ANSI, 1969; R1997). The underlying motivation for calculation of band-importance functions is that important speech information is contained in similar frequency bands, even for different speakers and across different listening conditions. An IBM that accounts for these inequivalent contributions of different frequencies to speech information would vary the LC across the frequency bands/filters. Processing would thus give preference to the retention of the "important" frequency bands and to the omission of the relatively "unimportant" bands.

2. RC Performance Function in Hearing Impaired Listeners

Another important future direction of this work is to test the effect of RC value in hearing-impaired (HI) participants. The performance of HI listeners with IBM-processed speech

as a function of different RC values is important for applying the IBM to technologies to aid those with hearing impairment. Healy et al. (2014) demonstrated that hearing-impaired listeners could benefit more than NH listeners from hearing IBM-processed relative to unprocessed/aCa/ phonemes in noise at RC = -4, -5, and -6 dB SNR. Hearing-impaired listeners are known to have a lower tolerance for noise than NH listeners. Therefore, testing their performance across different RC values may reveal an RC performance function that reflects a preference for discarding a greater proportion of the signal (and therefore more noise), and thus whose optimal values are slightly more positive than those currently proposed for NH listeners. Testing HI listeners with varying degrees of hearing loss and examining the correlation between degree of hearing loss and the RC value(s) at which optimal performance is obtained may reveal important information about the balance between noise tolerance and usefulness of speech information.

Because HI listeners are shown to perform more poorly on IBM-processed speech than NH listeners, testing HI listeners is another way to avoid a ceiling effect. Hearing-Impaired listeners may be tested on both /aCa/ phonemes as used in the current experiment and on sentences as used in previous studies to determine the correlation between optimal RC values (and noise tolerance) in different speech materials at ecologically valid SNRs.

3. Application to an Algorithm to Estimate the IBM

It is important to note that the required prior knowledge of the target and background signals used to identify speech- or noise-dominated T-F units is unavailable in a natural environment, and thus the IBM itself cannot be implemented in hearing aids or other technologies. However applications of the IBM are still possible through a recent technology. Healy and colleagues (Healy et al., 2013; Healy et al., 2014) describe a machine-learning

algorithm that estimates the ideal binary mask with no prior knowledge of the speech and noise. The effectiveness of the algorithm was striking. In fact, hearing-impaired participants listening to speech-plus-noise processed by the algorithm performed equivalently to *or even better* than normal-hearing listeners presented with the unprocessed speech-plus-noise. The comparison of these conditions simulated a real-life implementation of the algorithm in which only hearing-impaired listeners have access to the algorithm through hearing aids. These results represent a significant improvement in noise reduction, of a scale that has not been obtained in decades. Furthermore, this algorithm has many aspects that make future real-world implementation feasible.

The optimal RC values as determined by the current experiment (as well as other IBM improvements that may be facilitated by the determination of optimal RC values, such as the variable-LC mask) improve the performance of the IBM, and therefore have the potential to eventually advance the algorithmic estimation of the IBM. These developments could easily be incorporated into the Healy et al. (2013; 2014) algorithm to improve the performance of hearing-impaired individuals.

As mentioned above, previous generations of the algorithm estimated IBMs that had RC values of -4, -5, and -6 dB SNR (denoted by the box in Fig. 4). Selecting an RC value for the estimation goal that is closer to the center of the performance plateau may improve human listener performance with algorithmic estimations of the IBM more than performance with the IBM because it allows more room for error in the algorithm's decision to retain or discard a T-F unit.

It is hoped that the experiment described in this paper may lead to important future developments in the criterion to determine whether a T-F unit of a speech-noise mixture is retained or discarded in the Ideal Binary Mask, and in turn guide improvements in the algorithm to estimate the IBM. Such developments have the potential to reduce the struggles that hearing-impaired listeners face in noise, and therefore improve the quality of life of millions of people.

Acknowledgements

This work was supported in part by the 2014 Study of Language Variation Undergraduate Research Award and the Autumn 2015 College of Arts and Sciences Undergraduate Research Scholarship. Sarah Yoho and Dr. Eric Healy from the Speech and Hearing Science Department at The Ohio State University are thanked for their active roles in advising this project and in revising this document.

References

American National Standards Institute (1969). S3.5 (R1896), *American National Standard Methods for Calculation of the Speech Intelligibility Index* (Acoustical Society of America, New York).

American National Standards Institute (1997). S3.5 (R2007), *American National Standard Methods for the Calculation of the Speech Intelligibility Index* (Acoustical Society of America, New York).

Anzalone, M. C., Calandruccio, L., Doherty, K. A., & Carney, L. H. (2006). Determination of the potential benefit of time-frequency gain manipulation. *Ear Hear., 27,* 480-492.

Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. L. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *J. Acoust. Soc. Am., 120,* 4007-4018.

Cao, S., Li, L., & Wu, X. (2011). Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise. *J. Acoust. Soc. Am., 129,* 2227–2236.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am., 119,* 1562-1573.

Dillon, H. (2012). *Hearing aids* (2nd Ed). New York: Thieme.

Healy, E. W., Yoho, S. E., Wang, Y., & Wang, D. L. (2013). An algorithm to improve speech recognition in noise for hearing-impaired listeners. *J. Acoust. Soc. Am., 134,* 3029-3038.

Healy, E. W., Yoho, S. E., Wang, Y., Apoux, F., & Wang, D. L. (2014). Speech-cue transmission by an algorithm to increase recognition in noise for hearing-impaired listeners. *J. Acoust. Soc. Am., 136*, 3325-3336.

Hu, G. & Wang, D.L. (2001). Speech segregation based on pitch tracking and amplitude modulation. *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 79-82.

Kjems, U., Boldt, J. B., Pederson, M.S., Lunner, T., & Wang, D. L. (2009). Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *J. Acoust. Soc. Am., 126*, 1415-1426.

Li, N. & Loizou, P. C. (2008). Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *J. Acoust. Soc. Am., 123*, 1673-1682.

Simpson, S. & Cooke, M. P. (2005). Consonant identification in *N*-talker babble is a nonmonotonic function of *N*. *J. Acoust. Soc. Am., 118*, 2775–2778.

Sinex, D. G. (2013). Recognition of speech in noise after application of time-frequency masks: Dependence on frequency and threshold parameters. *J. Acoust. Soc. Am., 133*, 2390–2396.

Wang, D.L. (2005). On ideal binary mask as the computational goal of auditory scene analysis. In P. Divenyi (Ed.), *Speech separation by humans and machines* (pp. 181-197). Norwell, MA: Kluwer Academic.

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., & Lunner, T. (2008). Speech perception of noise with binary gains. *J. Acoust. Soc. Am., 124*, 2303–2307.

Wang, D. L., Kjems, U., Pederson, M. S., Boldt, J. B., & Lunner, T. (2009). Speech intelligibility in background noise with ideal binary time-frequency masking. *J. Acoust. Soc. Am., 125,* 2336-2347.

World Health Organization. (2014). Deafness and hearing loss, [Fact Sheet]. Retrieved August 2014 from http://www.who.int/mediacentre/factsheets/fs300/en/